Seven Ways to Write One AI Instruction Scored the Same. The Priciest Cost 45% More

A benchmark of 27 AI models by Paulo Teixeira graded 17,850 answers and found prompt format barely moves accuracy — but moves the token bill.

Paulo Teixeira, creator of the NTC prompt engineering framework, during an interview about the benchmark

MIAMI, FL, UNITED STATES, July 29, 2026 /EINPresswire.com/ -- Seven different ways of writing the same instruction were put to 27 artificial intelligence models across 17,850 graded answers. On accuracy, statistical testing could not tell them apart. On cost, the gap reached 45%.

The benchmark, concluded in July 2026, shifts the practical question from which notation makes a model smarter to which makes it cheaper. Six of the seven finished within three-quarters of a percentage point of one another, narrower than the study's own margin of error. Token consumption ran from 473 per prompt for Markdown in English to 686 for XML in Portuguese.

> “None of this came from talent or from luck. It came from repetition — from spending money and tokens getting it wrong, from testing even what the AI itself said would not work.”
>
> *Paulo Teixeira*

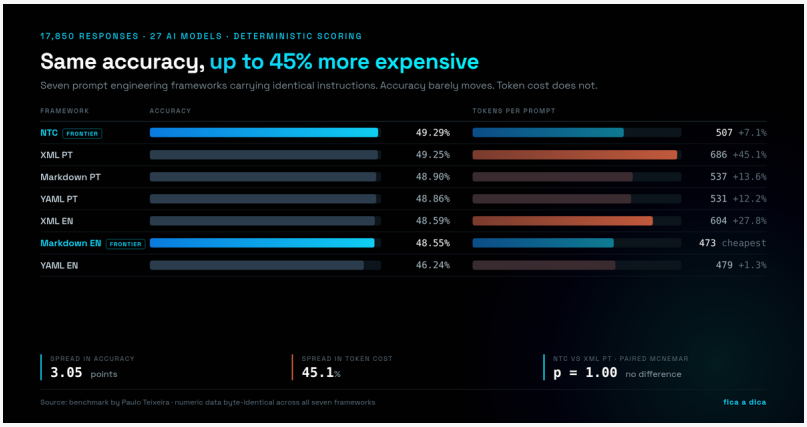
Counting tokens is deterministic. It carries no margin of error, and the difference repeats on every call — which is what turns a rounding error in a benchmark into a line item in an invoice. A system sending a million instructions a month pays the 45% every time, whether or not the model answered any better for it.

Only two formats held what the study calls the efficiency frontier — those no rival beats on accuracy and cost at once. Markdown in English held it by being the cheapest. The other was NTC prompt engineering, a proprietary notation with the field's highest accuracy at 49.29%.

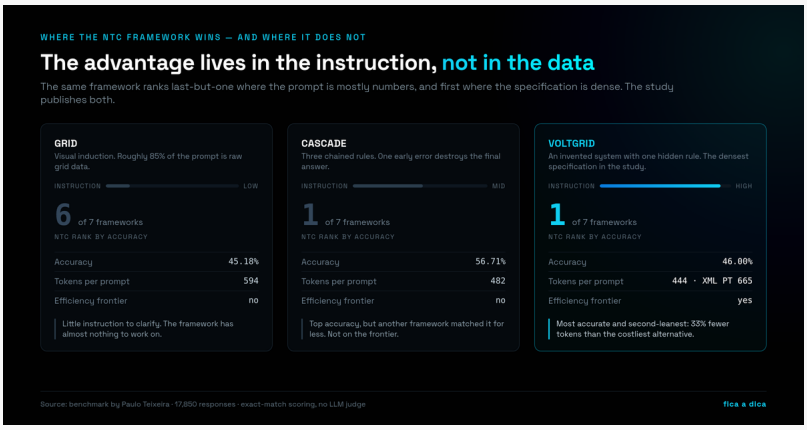
The NTC prompt engineering framework, created by Paulo Teixeira in 2025, was built to raise the clarity of an instruction rather than to compress it, with the reduction in tokens treated as a

consequence and not the objective.

Studies of this shape are usually run inside model laboratories and published with the parts that matter withheld. This one was run by one practitioner in João Pessoa, in the north-east of Brazil, and published with the cells attached — which, on the available evidence, makes it the largest open study of prompt engineering yet produced in the country. Teixeira attaches his own caveat to that claim: no central registry of such work exists, so it cannot be settled, only stated with the data in view.



Seven prompt engineering frameworks: accuracy spread of 3.05 points, token cost spread of 45.1%



The NTC framework ranks 6th of 7 where the prompt is mostly data, and 1st where the specification is dense

The test ran from 15 to 22 July 2026: 34 complete runs of 525 cells each, covering 27 distinct models from six providers, seven prompt formats and two languages, at a cost of roughly US$900 in API calls. Each of the 17,850 answers was checked against an answer key computed by code, character by character. No human judge, and no model grading another model's output — a practice common enough in prompt engineering research that its absence is worth naming. Either the model returned the exact sequence or it failed, and an unanswered cell scored zero.

The tasks came in three groups. GRID asked for visual induction in the style of ARC-AGI-2: infer a hidden rule from three or four worked examples, then apply it to a new input. CASCADE chained three transformations, each output feeding the next, so one early error destroyed the final answer. VOLTGRID was invented for the test — a documented system that existed nowhere on the internet before publication, with one rule deliberately omitted, leaving the model to infer it from execution traces and simulate the result. That design answers the quiet problem with public benchmarks: once a test is published, it drifts into the next generation's training data, and a model that has already read the answer key measures memory rather than thinking. A system that did not exist until the study went live cannot have been memorised by anything.

The difficulty landed where it was meant to. One puzzle defeated the entire field: no correct answer in 1,190 attempts. The easiest was solved 83.9% of the time, and the highest overall

score anywhere in the study was 77.5% — far enough from the ceiling that the test still has room to measure the models that come next.

The study's most unusual passage runs against the person who conducted it. NTC finished at 49.29% and XML in Portuguese at 49.25%, a difference of one correct answer in 2,550. Under a McNemar paired test, p = 1.00: no detectable difference. Against Markdown in Portuguese, p = 0.69. Also none. "A good study also shows where its own bet loses," Paulo Teixeira wrote in the case study published alongside the results.

The same file and method produced p-values on the order of 2×10^{-66} for other effects the work measured. That symmetry — one procedure for favourable and unfavourable findings alike — is what gives the remaining numbers their weight.

There is a trap inside that tie, and the NTC case study spells it out rather than hiding behind the average. Pooled across all 34 runs, the seven prompt engineering frameworks separate by just 3.06 points, which reads as though the choice barely matters. Inside a single run the median separation is 12.0 points, and in one run it reached 21.3. The aggregate is small because models disagree with each other: a framework that helps one model hurts another, and the gains cancel on their way into the average. As Paulo Teixeira puts it in the case study, the average does not measure the effect — it measures what is left after the effects cancel out. For a practitioner running one model rather than thirty-four, the number that applies is the one inside the run.

The cost picture is not uniformly flattering either. In the benchmark the NTC framework consumed 507 tokens per prompt against 473 for the leanest format: 7% more, not less. Roughly 85% of each test prompt is numeric data and only 15% instruction, and the framework's advantage lives in the instruction, so short prompts leave it little room to appear.

Where instruction dominates, the order inverts. In VOLTGRID, the densest specification in the set and the closest thing in it to a working system prompt, the NTC framework was both the most accurate format, at 46.00%, and the second-leanest, at 444 tokens against 665 for XML in Portuguese — a higher score on 33.2% fewer tokens than the priciest option. For anyone choosing a prompt engineering framework, that is the axis that decides: not which reasons better, but which says the same thing in less.

Outside the test set the margin widens. Paulo Teixeira measured 93 real production pairs — the same instruction before and after rewriting — with OpenAI's tokenizer. The set fell from 418,158 tokens to 182,121, a reduction of 56.45%, with a median of 55.72% per pair and a best case of 73.7%. One pair came out 3.3% larger, and that datapoint sits in the report. On the most bloated system prompts, around 40,000 tokens, he reports reductions of 70% to 80% — offered as production experience rather than controlled measurement, and not averaged into the rest.

"It works especially well on long system prompts, which were exactly my pain point," he said.

The distinction he draws is between compression and clarification. Compression pays for its saving with a piece of the quality. Clarification moves both in the same direction, because an instruction the model reads more easily is also a shorter one. Token reduction stops being the target and becomes the residue.

The work also declares its margin of error, which is rarer than it should be. With 525 cells per run, the 95% confidence interval on an overall score is roughly six points either way — wide enough to scramble the top of the ranking it publishes. Its top three finishers are statistically tied. The set separates bands of capability, not neighbours on a podium.

Read across all four variables it manipulated, the benchmark ranks the levers against one another — and that ordering is the part most likely to unsettle anyone shopping for a prompt engineering framework. Switching a model's extended reasoning on moved accuracy by 41.7 points — one economy model went from 6.5% to 48.2%. Changing the environment a model runs in, with the prompt untouched, moved it by as much as 20.6. Changing the framework moved it by a median of 12.0 within a run. Changing the language of the instruction moved it least. The ordering is uncomfortable for the field: the settings around the prompt matter more than the prompt's shape, and both matter more than which model is answering.

Language was the assumption the study set out to check, and the result refuses the usual advice. Portuguese finished at 49.01% against 47.79% for English — a gap of 1.22 points across 7,650 cells per language. Run by run, the scoreboard reads 16 wins for Portuguese, 16 for English and two ties. The recommendation to write prompts in English because models are trained mostly on English does not survive contact with a paired test. What survives is narrower and more useful: test the instruction language on the model and the task actually in front of you, because the effect exists but does not point one way.

The notation began in the second half of 2025, late at night, with ten versions of the same instruction set for an AI agent laid out for comparison. One looked strange, and the planning agent judged it would not work, since the model would first have to be taught the unfamiliar symbols. In testing, the strange one kept winning. Teixeira trusted the measurement over the opinion, took the instruction apart fragment by fragment and kept what survived. Asked whether it adapted some known method, the agent replied: "No, Paulo, NTC does not exist. It came out of these exhaustive tests of ours — you created this just now."

The problem was concrete rather than academic. System instructions above 40,000 tokens consumed the model's working memory before it began working, and in one bad week a Gemini API bill reached 10,000 Brazilian reais. The trouble was never the price per token, but that the instruction occupied the space where the reasoning was supposed to happen.

One finding from the same dataset is being held for separate publication: a top-tier model that refused to answer 43 of its 525 cells, blocked by a safety classifier before it ever saw the task. Everything else is on the table.

The material is open. The prompts for the six public formats, the rankings, and a file containing all 17,850 individual cells — each row recording run, task, format, language, repetition and whether the answer was correct — are published under a licence permitting use and adaptation, including commercially, with attribution. Because the file is paired, it supports the tests that require subject-by-subject comparison, so anyone can redo the arithmetic, including to contest it. Two things are withheld: the answer keys, so the benchmark stays usable against future models, and the NTC notation itself, which is proprietary and appears through its results rather than its contents.

Teixeira has more than 25 years in automation and SEO, with more than 25,000 orders delivered across more than 100 countries — figures from the Ana SEO Agency profile on the Fiverr platform, which carries the Vetted Pro badge and averages 4.9 stars across roughly 7,600 reviews. His AI usage runs to trillions of tokens over the career, more than 30 billion through Claude Code in a three-month span at the end of 2025, at a current pace of around 1.5 billion a day. Those figures measure scale of use, not competence — which is the argument for running a benchmark at all. Heavy use generates hypotheses; only measurement decides which survive. One that did not was his own, about his notation's advantage in accuracy.

"None of this came from talent or from luck," he said. "It came from repetition — from building the same project hundreds of times, from spending money and tokens getting it wrong, from testing even what the AI itself said would not work."

The NTC prompt engineering framework is not the only thing that came out of that practice. Paulo Teixeira is also the creator of Prompthen, a method for building and directing systems of AI agents in plain language, which gathers the working habits the benchmark measured one slice of.

What the study leaves behind is less a ranking than a reordering of the question. For two years the field has argued about which notation a model prefers, and the answer the cells return is that, on accuracy, it mostly does not have a preference worth paying for. The differences that survive a paired test sit elsewhere: in whether reasoning is switched on, in the environment the model runs inside, and — once accuracy has tied — in what the instruction costs to send, on every call, for as long as the system runs.

About the NTC prompt engineering framework

NTC is a prompt engineering framework created by Paulo Teixeira in 2025, during the optimization of system prompts in live operation with AI agents. Its principle is to increase the clarity of an instruction for the model; token reduction is the consequence, not the objective. The NTC framework has passed two independent validations: a comprehension study in October 2025, covering 36 tests across 6 models from 3 companies, and the benchmark concluded in July 2026, in which 17,850 answers were graded against a computed key across 27 models from 6

providers, with prompts, rankings and cell-by-cell results published openly. The full case study is published at https://ficaadica.com.br/novidades/ntc-modelos-performam-melhor/ and the dataset at https://github.com/pauloctxr/prompthen-bench. The NTC prompt engineering framework forms part of Prompthen (https://prompthen.ai), a method for building and directing systems of AI agents in plain language, also created by Paulo Teixeira. He has more than 25 years in automation and SEO, with more than 25,000 orders delivered across more than 100 countries according to the Ana SEO Agency profile on Fiverr, and shares his practice on the YouTube channel "Fica a Dica com Paulo Teixeira".

Paulo Teixeira
SEO Experts
email us here
Visit us on social media:
LinkedIn
Instagram
YouTube

This press release can be viewed online at: https://www.einpresswire.com/article/930102565

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.
© 1995-2026 Newsmatics Inc. All Right Reserved.